Disclosure Limitation and Confidentiality Protection in Linked Data

by

John M. Abowd
U.S. Census Bureau and Cornell University

Ian M. Schmutte
University of Georgia

Lars Vilhuber
Cornell University

CES 18-07 **January, 2018**

Abstract

Confidentiality protection for linked administrative data is a combination of access modalities and statistical disclosure limitation. We review traditional statistical disclosure limitation methods and newer methods based on synthetic data, input noise infusion and formal privacy. We discuss how these methods are integrated with access modalities by providing three detailed examples. The first example is the linkages in the Health and Retirement Study to Social Security Administration data. The second example is the linkage of the Survey of Income and Program Participation to administrative data from the Internal Revenue Service and the Social Security Administration. The third example is the Longitudinal Employer-Household Dynamics data, which links state unemployment insurance records for workers and firms to a wide variety of censuses and surveys at the U.S. Census Bureau. For examples, we discuss access modalities, disclosure limitation methods, the effectiveness of those methods, and the resulting analytical validity. The final sections discuss recent advances in access modalities for linked administrative data.

[*] The authors acknowledge the support of a grant from the Alfred P. Sloan Foundation (G-2015-13903) and NSF Grants SES-1131848, BCS-0941226, TC-1012593. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the National Science Foundation, or the Sloan Foundation. All results presented in this work stem from previously released work, were used by permission, and were previously reviewed to ensure that no confidential information is disclosed.

John M. Abowd is the Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, the Edmund Ezra Day Professor of Economics, Professor of Statistics and Information Science, and the Director of the Labor Dynamics Institute (LDI) at Cornell University, Ithaca, NY, USA. https://johnabowd.com. Ian M. Schmutte is Associate Professor of Economics at the University of Georgia, Athens, GA, USA. http://ianschmutte.org. Lars Vilhuber is Senior Research Associate in the Department of Economics and Executive Director of Labor Dynamics Institute (LDI) at Cornell University, Ithaca, NY, USA. https://lars.vilhuber.com.

Introduction[1]

The use of administrative data has long been a part of the procedures at national statistical offices (NSOs), as evidenced by the various chapters in this book. The censuses and surveys conducted by NSOs may use sampling frames built at least partially from administrative data. For instance, the U.S. Census Bureau has used a business register - a list of all domestic businesses - derived from administrative tax filings since at least 1968. This register is the frame for its quinquennial censuses and annual surveys of business activity (DeSalvo et al. 2016). It is also used to link businesses across surveys, to link surveyed businesses to other administrative record data, and as a direct source of statistical information on the levels and growth of business activity, published as the County Business Patterns (CBP) and Business Dynamics Statistics (BDS).[2] Similar examples can be found in most countries that maintain some kind of registry for their businesses. In many countries, similar centrally maintained registers are used as frames for censuses and surveys of a country's inhabitants and workers. Chapter 17 illustrates the Swedish approach to this problem for a national population census.[3] The Institute for Employment Research (IAB), the research institute of the German Employment Agency, uses social security notifications filed by firms, and data generated from the administration of its mandated programs, to sample firms and workers. McMaster University and later Statistics Canada used administrative job termination notifications ("record of employment") filed by employers to survey departing employees for the Canadian Out-of-Employment Panel (COEP) (Browning et al. 1995). Other uses of administrative data in NSOs include linkage for quality purposes (Chapters 8, 14 and 15), and data augmentation (Chapter 12 for the National Center for Health Statistics approach)

In addition, the increasing computerization of administrative records, has facilitated more extensive linking of previously disconnected administrative databases, to create more comprehensive and extensive information. Methods to link databases within administrative units based on common identifiers are easy to implement (see Chapter 9 for more details). In the United States, which does not have a legal national identifier or ID document, the increased use of the social security number (SSN) has facilitated linkage of government databases and among commercial data providers. In many European countries, individuals have national identifiers,

---

[1] John M. Abowd is the Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, the Edmund Ezra Day Professor of Economics, Professor of Statistics and Information Science, and the Director of the Labor Dynamics Institute (LDI) at Cornell University, Ithaca, NY, USA. https://johnabowd.com Ian M. Schmutte is Associate Professor of Economics at the University of Georgia, Athens, GA, USA. http://ianschmutte.org. Lars Vilhuber is Senior Research Associate in the Department of Economics and Executive Director of Labor Dynamics Institute (LDI) at Cornell University, Ithaca, NY, USA. https://lars.vilhuber.com. The authors acknowledge the support of a grant from the Alfred P. Sloan Foundation (G-2015-13903), NSF Grants SES-1131848, BCS-0941226, TC-1012593. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the National Science Foundation, or the Sloan Foundation. All results presented in this work stem from previously released work, were used by permission, and were previously reviewed to ensure that no confidential information is disclosed.

[2] See www.census.gov/programs-surveys/cbp.html and www.census.gov/ces/dataproducts/bds/.

[3] In the United States, a combination of multiple lists and input by regional and subject matter experts is used to compile the frame for the Census of Population and Housing.

and efforts are underway to allow for cross-border linkages within the European Union, in order to improve statistics on the workforce and the businesses of the common economic area created by what is now called the European Union. However, even when common identifiers are not available, linkage is possible (see Chapter 15).

The result has been that data on individuals, households and business has become richer, collected from an increasing variety of sources, both as designed surveys and censuses, as well as organically created "administrative" data. The desire to allow policy makers and researchers to leverage the rich linked data has been held back, however, by the concerns of citizens and businesses about privacy. In the 1960s in the United States, researchers had proposed a "National Data Bank" with the goal of combining survey and administrative data for use by researchers. Congress held hearings on the matter, and ultimately the project did not go forward (Kraus 2013). Instead, and partially as a consequence, privacy laws were formalized in the 1970s. The U.S. "Privacy Act" (Public Law 93-579, 5 U.S.C. § 552a), passed in 1974, specifically prohibited "matching" programs, linking data from different agencies. More recently, the 2016 Australian Census elicited substantial controversy when the Australian Bureau of Statistics (ABS) decided to keep identifiable data collected through the census for a substantially longer time period, with the explicit goal of enabling linkages between the census and administrative data, as well as linkages across historical censuses (Australian Bureau of Statistics 2015; Karp 2016).

Subsequent decades saw a decline in public availability of highly detailed microdata on people, households, and firms, and the emergence of new access mechanisms and data protection algorithms. This chapter will provide an overview of the methods that have been developed and implemented to safeguard privacy, while providing researchers the means to draw valid conclusions from protected data. The protection mechanisms we will describe are both physical and statistical (or algorithmic), but exist because of the need to balance the privacy of the respondents, including the confidentiality protection their data receive, with society's need and desire for ever more detailed, timely, and accurate statistics.

Paradigms of protection

There are no methods for disclosure limitation and confidentiality protection specifically designed for linked data. Protecting data constructed by linking administrative records, survey responses, and "found" transaction records relies on the same methods as might be applied to each source individually. It is the richness inherent in the linkages, and in the administrative information available to some potential intruders, that poses novel challenges.

Statistical confidentiality can be viewed as "a body of principles, concepts, and procedures that permit confidentiality to be afforded to data, while still permitting its use of for statistical purposes" (Duncan et al. 2011, p.2). In order to protect the confidentiality of the data they collect, NSOs and survey organizations (henceforth referred to generically as data custodians) employ many methods. Very often, data are released to the public as tabular summaries. Many of the protection mechanisms in use today evolved to protect published *tables* against

disclosure. Generically, the idea is to limit the publication of cells with “too few” respondents, where the notion of “too few” is assessed heuristically.

We will not provide a detailed history or taxonomy of statistical disclosure limitation (SDL) and formal privacy models, instead referring the reader to other publications on the topic (Duncan et al. 2011; Dwork & Roth 2014; FCSM 2005). We do need to set up the problem, which we will do by reviewing suppression, coarsening, swapping, and noise infusion (input and output). These are widely used techniques and the main issues that arise in applications to linked data can be understood with reference to these methods.

Suppression is widely used to protect published tables against statistical disclosure. Suppression describes the removal of sub-tables, cells or items in a cell from a published collection of tables if the item’s publication would pose a high risk of disclosure. This method attempts to forge a middle ground between the users of tabular summaries, who want increasingly detailed disaggregation, and publication rules based on cell count thresholds. The Bureau of Labor Statistics (BLS) uses suppression as its primary SDL technique for data releases based on business establishment censuses and surveys. From the outset, it was understood that primary suppression -- not publishing easily identified data items -- didn't protect anything if the agency published the rest of the data, including summary statistics. Users could infer the missing items from what was published (Fellegi 1972). The BLS, and other agencies that rely on suppression, make “complementary suppressions” to reduce the probability that a user can infer the sensitive items from the published data (Holan et al. 2010). But there is no optimal complementary suppression technology - there are usually multiple complementary suppression strategies that achieve the same protection.

Researchers, however, are not indifferent among these strategies. A researcher who needs detailed geographic variation will benefit from data in which the complementary suppressions are based on removing detailed industries. A researcher who needs detailed industry variation will prefer data with complementary suppression based on geography. Ultimately, the committee that chooses the complementary suppression strategy will determine which research uses are possible and which are ruled out.

But the problem is deeper than this: suppression is a very ineffective SDL technique. Researchers working with the cooperation of the BLS have shown that the suppression strategy used in major BLS business data publications provides almost no protection if it is applied, as is currently the case, to each data release separately (Holan et al. 2010). Some agencies may use cumulative suppression strategies in their sequential data releases. In this case, once an item has been designated for either primary or complementary suppression, it would disappear from the release tables until the entire product is redesigned.

Many social scientists believe that suppression can be complemented by restricted access agreements that allow the researcher to use all of the confidential data but limit what can be published from the analysis. Such a strategy is not a complete solution because SDL must still

be applied to output of the analysis, which quickly brings the problem of what output to suppress back to the forefront.

Custom tabulations and data enclaves Another traditional response by data custodians to the demand by researchers for more extensive and detailed summaries of confidential data, was to create a custom tabulation, a table not previously published, but generated by data custodian staff with access rights to the confidential data, and typically subject to the same suppression rules. As these requests increased, the tabulation and analysis work was offloaded onto researchers by providing them with access to protected microdata. This approach has expanded rapidly in the last two decades, and is widely used around the world. We discuss it in more detail later in this chapter.

Coarsening is a method for protecting data that involves mapping confidential values into broader categories. The simplest method is a histogram, which maps values into (fixed) intervals. Intuitively, the broader the interval, the more protection is provided.

Sampling is a protection mechanism that can be applied either at the collection stage or at the data publication stage. At the collection stage, it is a natural part of conducting surveys. In combination with coarsening and the use of statistical weights, the basic idea is simple: if a table cell is based on only a few sampled individuals which collectively represent the underlying population, then statistical inference will not reveal the attributes of any particular individual with any precision, as long as the identity of the sampled individuals is not revealed. Both coarsening and sampling underlie the release of public use microdata samples.

Input noise infusion

Protection mechanisms for microdata are often similar in spirit, though not in their details, to the methods employed for tabular data. Consider coarsening, in which the more detailed response to a question (say, about income), is classified into a much smaller set of bins (for instance, income categories such as “[10,000; 25,000]”). In fact, many tables can be viewed as a coarsening of the underlying microdata, with a subsequent count of the coarsened cases.

Many microdata methods are based on **input noise infusion**: distorting the value of some or all of the inputs before any publication data are built. The Census Bureau uses this technique before building publication tables for many of its business establishment products and in the American Community Survey (ACS) publications, and we discuss it in more detail for one of those data products later in this chapter. The noise infusion parameters can be set such that all of the published statistics are formally unbiased--the expected value of the published statistic equals the value of the confidential statistic with respect to the probability distribution of the infused noise--or nearly so. Hence, the disclosure risk and data quality can be conveniently summarized by two parameters: one measuring the absolute distortion in the data inputs and the other measuring the mean squared error of publication statistics (either overall for censuses or relative to the undistorted survey estimates).

From the viewpoint of empirical social sciences, however, all input distortion systems with the same risk-quality parameters are not equivalent. In a regression discontinuity design, for example, there will now be a window around the break point in the running variable that reflects the uncertainty associated with the noise infusion. If the effect is not large enough, it will be swamped by noise even though all the inputs to the analysis are unbiased, or nearly so. Once again, using the unmodified confidential data via a restricted access agreement doesn't completely solve the problem because once the noisy data have been published, the agency has to consider the consequences of allowing the publication of a clean regression discontinuity design estimate where the plot of the unprotected outcomes vs. the running variable can be compared to the similar plot produced from the public noisy data.

An even more invasive input noise technique is **data swapping**. Sensitive data records (usually households) are identified based on *a priori* criteria. Then, sensitive records are compared to "nearby" records on the basis of a few variables. If there is a match, the values of some or all of the other variables are swapped (usually the geographic identifiers, thus effectively relocating the records in each other's location). The formal theory of data swapping was developed shortly after the theory of primary/complementary suppression (Dalenius & Reiss 1982, first presented at American Statistical Association (ASA) Meetings in 1978). Basically, the marginal distribution of the variables used to match the records is preserved at the cost of all joint and conditional distributions involving the swapped variables. In general, very little is published about the swapping rates, the matching variables, or the definition of "nearby," making analysis of the effects of this protection method very difficult. Furthermore, even arrangements that permit restricted access to the confidential files still require the use of the swapped data. Some providers destroy the unswapped data. Data swapping is used by the Census Bureau, National Center for Health Statistics (NCHS), and many other agencies (FCSM 2005). The Census Bureau does not allow analysis of the unswapped decennial and ACS data except under extraordinary circumstances that usually involve the preparation of linked data from outside sources then re-imposition of the original swap (so the records acquire the correct linked information, but the geographies are swapped according to the original algorithm before any analysis is performed). NCHS allows the use of unswapped data in its restricted access environment but prohibits publication of most subnational geographies when the research is published.

The basic problem for empirical social scientists is that agencies must have a general purpose data publication strategy in order to provide the public good that is the reason for incurring the cost of data collection in the first place. But this publication strategy inherently advantages certain analyses over others. Statisticians and computer scientists have developed two related ways to address this problem: synthetic data combined with validation servers and privacy-protected query systems. Statisticians define "synthetic data" as samples from the joint probability distribution of the confidential data that are released for analysis. After the researcher analyzes the synthetic data, the validation server is used to repeat some or all of the analyses on the underlying confidential data. Conventional SDL methods are used to protect the statistics released from the validation server.

Formal privacy models

Computer scientists define a privacy-protected query system as one in which all analyses of the confidential data are passed through a noise-infusion filter before they are published. Some of these systems use input noise infusion--the confidential data are permanently altered at the record level, and then all analyses are done on the protected data. Other formally private systems apply output noise infusion to the results of statistical analyses before they are released.

All formal privacy models define a cumulative, global privacy loss associated with all of the publications released from a given confidential database. This is called the total privacy-loss budget. The budget can then be allocated to each of the released queries. Once the budget is exhausted, no more analyses can be conducted. The researcher must decide how much of the privacy-loss budget to spend on each query--producing noisy answers to many queries or sharp answers to a few. The agency must decide the total privacy-loss budget for all queries and how to allocate it among competing potential users.

An increasing number of modern SDL and formal privacy procedures replace methods like deterministic suppression and targeted random swapping with some form of noisy query system. Over the last decade these approaches have moved to the forefront because they provide the agency with a formal method of quantifying the global disclosure risk in the output and of evaluating the data quality along dimensions that are broadly relevant.

Relatively recently, formal privacy models have emerged from the literature on database security and cryptography. In formal privacy models, the data are distorted by a randomized mechanism prior to publication. The goal is to explicitly characterize, given a particular mechanism, how much private information is leaked to data users.

Differential privacy is a particularly prominent and useful approach to characterizing formal privacy guarantees. Briefly, a formal privacy mechanism that grants ε-differential privacy places an upper bound, parameterized by ε, on the ability of a user to infer from the published output whether any specific data item, or response, was in the original, confidential data (see Dwork & Roth 2014 for an in depth discussion).

Formal privacy models are very intriguing because they solve two key challenges for disclosure limitation. First, formal privacy models by definition provide provable guarantees on how much privacy is lost, in a probabilistic sense, in any given data publication. Second, the privacy guarantee does not require that the implementation details, specifically the parameter ε, be kept secret. This allows researchers using data published under formal privacy models to conduct fully *SDL-aware* analysis. This is not the case with many traditional disclosure limitation methods which require that key parameters, such as the swap rate, suppression rate, or variance of noise, not be made available to data users (Abowd & Schmutte 2015).

Confidentiality protection in linked data: Examples

To illustrate the application of new disclosure avoidance techniques, we describe three examples of linked data and the means by which confidentiality protection is applied to each. First, the **Health and Retirement Study (HRS)** links extensive survey information to respondents' administrative data from the Social Security Administration (SSA) and the Center for Medicare and Medicaid Services (CMS). To protect confidentiality in the linked HRS-SSA data, its data custodians use a combination of restrictive licensing agreements, physical security, and restrictions on model output. Our second example is the Census Bureau's Survey of Income and Program Participation (SIPP), which has also been linked to earnings data from the Internal Revenue Service (IRS) and benefit data from the SSA. Census makes the linked data available to researchers as the **SIPP Synthetic Beta File**. Researchers can directly access synthetic data via a restricted server and, once their analysis is ready, request output based on the original harmonized confidential data via a validation server. Finally, the **Longitudinal Employer-Household Dynamics Program** (LEHD) at the Census Bureau links data provided by 51 state administrations to data from federal agencies and surveys and censuses on businesses, households, and people conducted by the Census Bureau. Tabular summaries of LEHD are published with greater detail than most business and demographic data. The LEHD is also accessible in restricted enclaves, but there are also restrictions on the output researchers can release. There are many other linked data sources. These three are each innovative in some fashion, and allow us to illustrate the issues faced when devising disclosure avoidance methods for linked data.

HRS-SSA

Data description

The Health and Retirement Study (HRS) is conducted by the Institute for Social Research at the University of Michigan. Data collection was launched in 1992 and has re-interviewed the original sample of respondents every two years since then. New cohorts and sample refreshment have made the HRS one of the largest representative longitudinal samples of Americans over 50, with over 26,000 respondents in a given wave (Sonnega & Weir 2014). In 2006, the HRS started collecting measures of physical function, biomarkers, and DNA samples. The collection of these additional sensitive attributes reinforce confidentiality concerns.

Linkages to other data

The HRS team requests permission from respondents to link their survey responses to other data resources, as described below. For consenting respondents, HRS data are linked at the individual level to administrative records from Social Security and Medicare claims, thus allowing for detailed characterizations of income and wealth over time.

The Centers for Medicare and Medicaid Services (CMS) maintain claims records for the medical services received by essentially all Americans age 65 and older and those less than 65 years

who receive Medicare benefits. These records include comprehensive information about hospital stays, outpatient services, physician services, home health care, and hospice care. When linked to the HRS interview data, this supplementary information provides far more detail on the health circumstances and medical treatments received by HRS participants than would otherwise be available.

Data from HRS interviews are also linked to information about respondents' employers. This improves information on employer-provided benefits, including pensions. While most pension-eligible workers have some idea of the benefits available through their pension plans, they generally are not knowledgeable about detailed provisions of the plans. By linking HRS interview data with detailed information on pension plans, researchers can better understand the contribution of the pension to economic circumstances and the effects of the pension structure on work and retirement decisions.

Sidebar: Select administrative data linked to HRS

CMS

HRS Medicare Claims and Summary Data (2012) Cross-Reference

SSA Administrative Data

Cross-Wave Social Security Weights

Supplemental Security Income [Respondent; Deceased Spouse]

Deceased Spouse Cross-Year Benefits

Respondent Cross-Year Benefits

Respondent Cross-Year Summary Earnings

Respondent Cross-Year Detail Earnings

Deceased Spouse Cross-Year Summary Earnings

Deceased Spouse Cross-Year Detail Earnings

Form 831 Disability Records [Respondent; Deceased Spouse]

Source: http://hrsonline.isr.umich.edu/index.php?p=reslis, as of August 2016

HRS data are also linked at the individual level to administrative records from Social Security and Medicare, Veteran's Administration, the National Death Index, and employer-provided pension plan information (Sonnega & Weir 2014).

Disclosure avoidance methods

To ensure privacy and confidentiality, all study participants' names, addresses, and contact information are maintained in a secure control file.[4] Anyone with access to identifying information must sign a pledge of confidentiality. The survey data are only released to the research community after undergoing a rigorous process to remove or mask any identifying information. First a set of sensitive variables (such as state of residence or specific occupation) are suppressed or masked. Next, the remaining variables are tested for any possible identifying content. When testing is complete, the data files are subject to final review and approval by the HRS Data Release Protocol Committee. Data ready for public use are made available to qualified researchers via a secure website. Registration is required of all researchers before downloading files for analyses. In addition, use of linked data from other sources, such as Social Security or Medicare records, is strictly controlled under special agreements with specially approved researchers operating in secure computing environments that are periodically audited for compliance.

Additional protections involve distortion of the microdata prior to dissemination to researchers. Earnings and benefits variables such as those from SSA in the HRS are rounded or top coded (Deang & Davies 2009). Similarly, geographic classifications are limited to broad levels of aggregation (for example, census divisions instead of states, or states instead of counties). The HRS uses licensing as its primary method of giving access to restricted files. A license can be secured only after meeting a stringent set of criteria that leads to a contractual agreement between the HRS, the researcher, and the researcher's employer. The license enables the user to receive restricted files and use them at the researcher's own institutional facility.

SIPP-SSA-IRS (SSB)

Data description

The SIPP/SSA/IRS Public Use File, known as the SIPP Synthetic Beta File or SSB, combines variables from the Census Bureau's SIPP, the IRS individual lifetime earnings data, and the SSA individual benefit data. Aimed at a user community that was primarily interested in national retirement and disability programs, the selection of variables for the proposed SIPP/SSA/IRS-PUF focused on the critical demographic data to be supplied from the SIPP, earnings histories going back to 1937 from the IRS data maintained at SSA, and benefit data from SSA's master beneficiary records, linked using respondents' Social Security Numbers. After attempting to determine the feasibility of adding a limited number of variables from the SIPP directly to the linked earnings and benefit data, it was decided that the set of variables that could be added without compromising the confidentiality protection of the existing SIPP public use files was so limited that alternative methods had to be used to create a useful new file.

[4] (National Institute on Aging and the National Institutes of Health n.d.)

The technique adopted is called *partially synthetic data* with *multiple imputation of missing items*. As the term is used in this chapter, "partially synthetic data" means the person-level records are released containing some variables from the actual responses and other variables where the actual responses have been replaced by values sampled from the posterior predictive distribution for that record, conditional on all of the confidential data. From 2003 until 2015, seven preliminary versions of the SSB were produced. In this chapter, we will focus on the protections that pertain to the linked nature of the data. The interested reader is referred to (Abowd et al. 2006) for details on data sources, imputation, and linkage. The analysis here is for the SSB version 4. Since version 4, two additional versions have been released with slightly different structure.[5] Subsequent versions are well-illustrated by the extensive analysis described here.

Disclosure avoidance methods

The existence of SIPP public use files poses a key challenge for disclosure avoidance. To protect the confidentiality of survey respondents, it was deemed necessary to prevent re-identification of a record that appears in the synthetic data against the existing SIPP public use files. Hence, all information regarding the dating of variables whose source was a SIPP response, and not administrative data, has to be made consistent across individuals regardless of the panel and wave from which the response was taken. The public use file contains several variables that were never missing and are not synthesized. These variables are: gender, marital status, spouse's gender, initial type of Social Security benefits, type of Social Security benefits in 2000, and the same benefit type variables for the spouse. All other variables in the SSB v4 were synthesized.

The model first imputes any missing data, then synthesizes the completed data (Reiter 2004). For each iteration of the missing data imputation phase and again during the synthesis phase, a joint posterior predictive distribution for all of the required variables is estimated according to the following protocol. At each node of the parent/child tree, a statistical model is estimated for each of the variables at the same level. The statistical model is a Bayesian bootstrap, logistic regression, or linear regression (possibly with transformed inputs). The missing data phase included nine iterations of estimation. The synthetic data phase occurred on the tenth iteration. Four missing data implicates were created. These constitute the completed data files that are the inputs to the synthesis phase. Four synthetic implicates were created for each missing data implicate, for a total of 16 synthetic implicates on the released file. Because copying the final weight to each implicate of the synthetic data would have provided an additional unsynthesized variable with 55,552 distinct values, the disclosure risk associated with the weight variable had to be addressed. A synthetic weight using a posterior predictive distribution based on the Multinomial/Dirichlet natural conjugate likelihood and prior was created.

[5] The latest version as of this writing is version 6.0.2 (U.S. Census Bureau 2015).

Disclosure avoidance assessment

The link of administrative earnings, benefits and SIPP data adds a significant amount of information to an already very detailed survey and could pose potential disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information found in the summary and detailed earnings histories used to create the longitudinal earnings variables.

The Census Bureau Disclosure Review Board at the time of release used two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. Second, the ratio of true matches to the total number of matches (true and false) should be close to one-half.

The disclosure avoidance analysis (Abowd et al. 2006) uses the principle that a potential intruder would first try to re-identify the source record for a given synthetic data observation in the existing SIPP public use files. Two distinct matching exercises - one *probabilistic* (Fellegi & Sunter 1969), one *distance-based* (Torra et al. 2006) - between the synthetic data and the harmonized confidential data were conducted.[6] The harmonized confidential data -- actual values of the data items as released in the original SIPP public use files -- are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the harmonized confidential data and the synthetic data represent potential disclosure risks. In practice, the intruder would also need to make another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, etc. The results from the experiments are conservative estimates of re-identification risk. For the probabilistic matching, the assessment matched synthetic and confidential files exactly on the unsynthesized variables of gender and marital status, and success of the matching exercise is assessed using a person identifier which is not, in fact, available in the released version of the synthetic data. Without the personid, an intruder would have to compare many more record pairs to find true matches, would not find any more true matches (the true match is guaranteed to be in the blocks being compared), and would almost certainly find more false matches. In fact, the records that can be re-identified represent only a very small proportion (less than three percent) of candidate records, and correct re-identifications are swamped by a sea of false re-identifications (Abowd et al. 2006, p.6).

In distance-based matching, records between the harmonized confidential and synthetic data are blocked in a similar way, and distances (or similarity scores) are computed for a given confidential record and every synthetic record within a block. The three closest records are

[6] In much of the documentation for the SSB, the internal confidential files, harmonized across the SIPP panets and waves, and completed using the multiple imputation procedures that produced the four implicates at the root of the synthesis for confidentiality protection, are called the "Gold Standard" files. This nomenclature means that these are the files that would be provided to a researcher in the Census Bureau's restricted access environment (FSRDC). Chapter 9 in this volume discusses linking methodologies.

declared matches, and the personid again checked to verify how often a true match is obtained. A putative intruder who treated the closest record as a match would correctly link about 1 percent of all synthetic records, and less than 3% in the worst-case sub-group (Abowd et al. 2006, p.8).

Analytical validity assessment

Although synthetic data are designed to solve a confidentiality protection problem, the success of this solution is measured by both the degree of protection provided and the user's ability to reliably estimate scientifically interesting quantities. The latter property of the synthetic data is known as analytical (or statistical) validity. Analytical validity exists when, at a minimum, estimands can be estimated without bias and their confidence intervals (or the nominal level of significance for hypothesis tests) can be stated accurately (Rubin 1987). To verify analytical validity, the confidence intervals surrounding the point estimates obtained from confidential and synthetic data should completely overlap (Reiter et al. 2009), presumably with the synthetic confidence interval being slightly larger because of the increased variation arising from the synthesis. When these results obtain, inferences drawn about the coefficients will be consistent whether one uses synthetic or completed data.

As an example, Figure 1[7] compares employment rates for black men and women.[8] The estimated percentage of individuals who worked in a given year is very close, on average, for both groups and across all the years. The confidence intervals overlap for most years. Similar results obtain for whites, as well as for average earnings for all the groups (see Abowd et al. 2006 for additional statistics).

Similar comparisons can be made for model-based results. Figure 2 reports coefficients from regressions of the log of total earnings[9] in the year 2000 on various explanatory variables, by sex for blacks. The closest correspondence between the synthetic and completed regression coefficients is in the education variables, which always have the same sign and generally have significant overlap in the confidence intervals. The exceptions for overlapping confidence intervals are usually the graduate degree indicator. Other SIPP demographic variables are not as consistently similar between the synthetic and completed data, but generally, there is some degree of overlap between confidence intervals, suggesting that the synthetic data are not producing estimates that are entirely different from the completed data.

[7] All data underlying these and other graphs, with one exception, have been made available at Vilhuber et al (2017).
[8] Strictly speaking, the statistics are for indicators of positive FICA-covered earnings on the SSA-provided Summary Earnings Records (SER) for individuals who became OASDI beneficiaries during the time period covered by these data i.e., had date of initial entitlement between 1951 and 2002 (Abowd et al. 2006, Table 3-9).
[9] Strictly speaking, sum of deferred and non-deferred earnings at FICA and non-FICA jobs on the Detailed Earnings Records (DER) (Abowd et al. 2006, Table 41, 43).

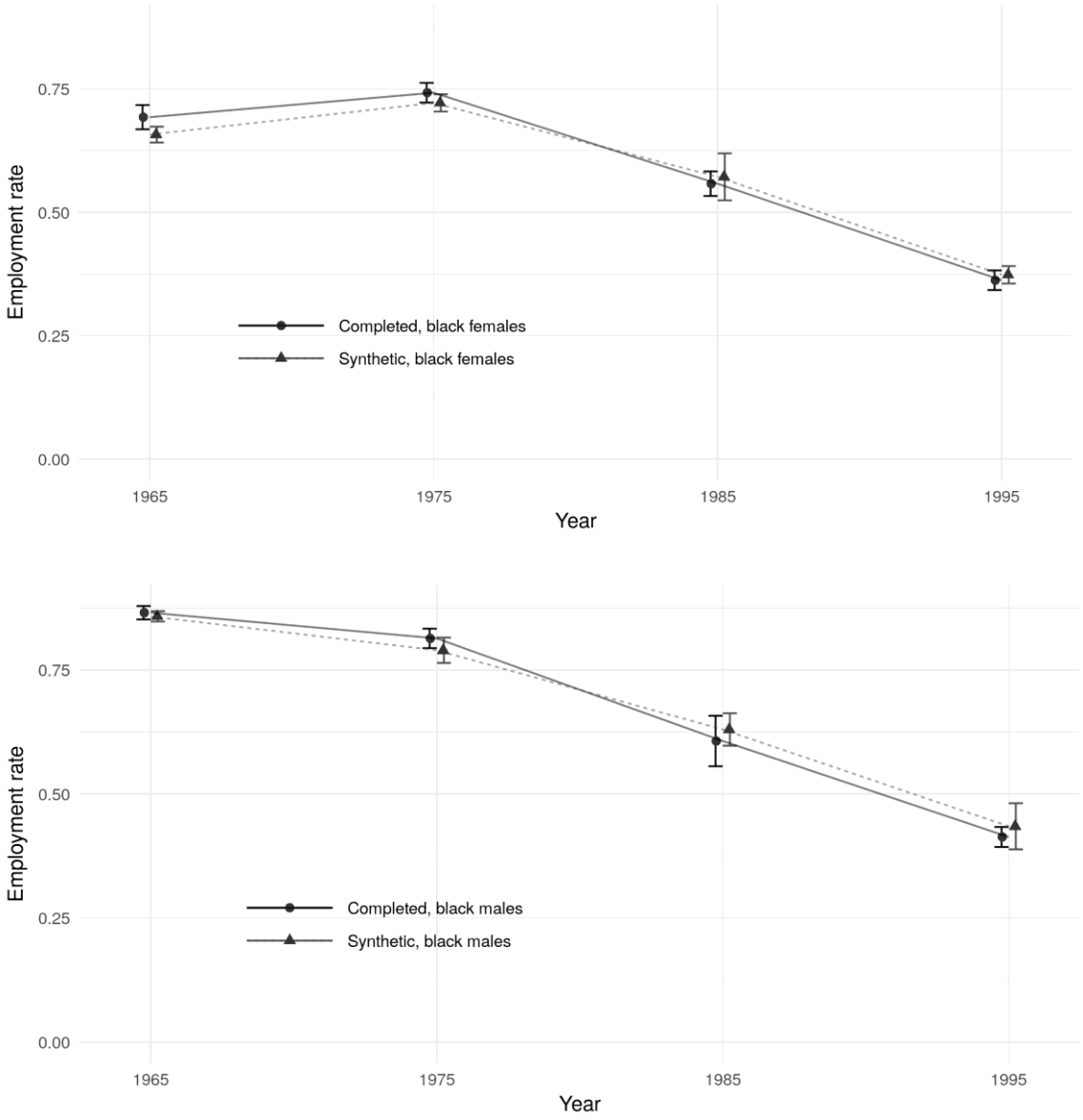

Figure 1: Comparison of employment rate, estimated from completed (confidential) data and from synthetic data, for black men and women. (Source: Abowd et al. 2006, Tables 3, 5, 7, 9)

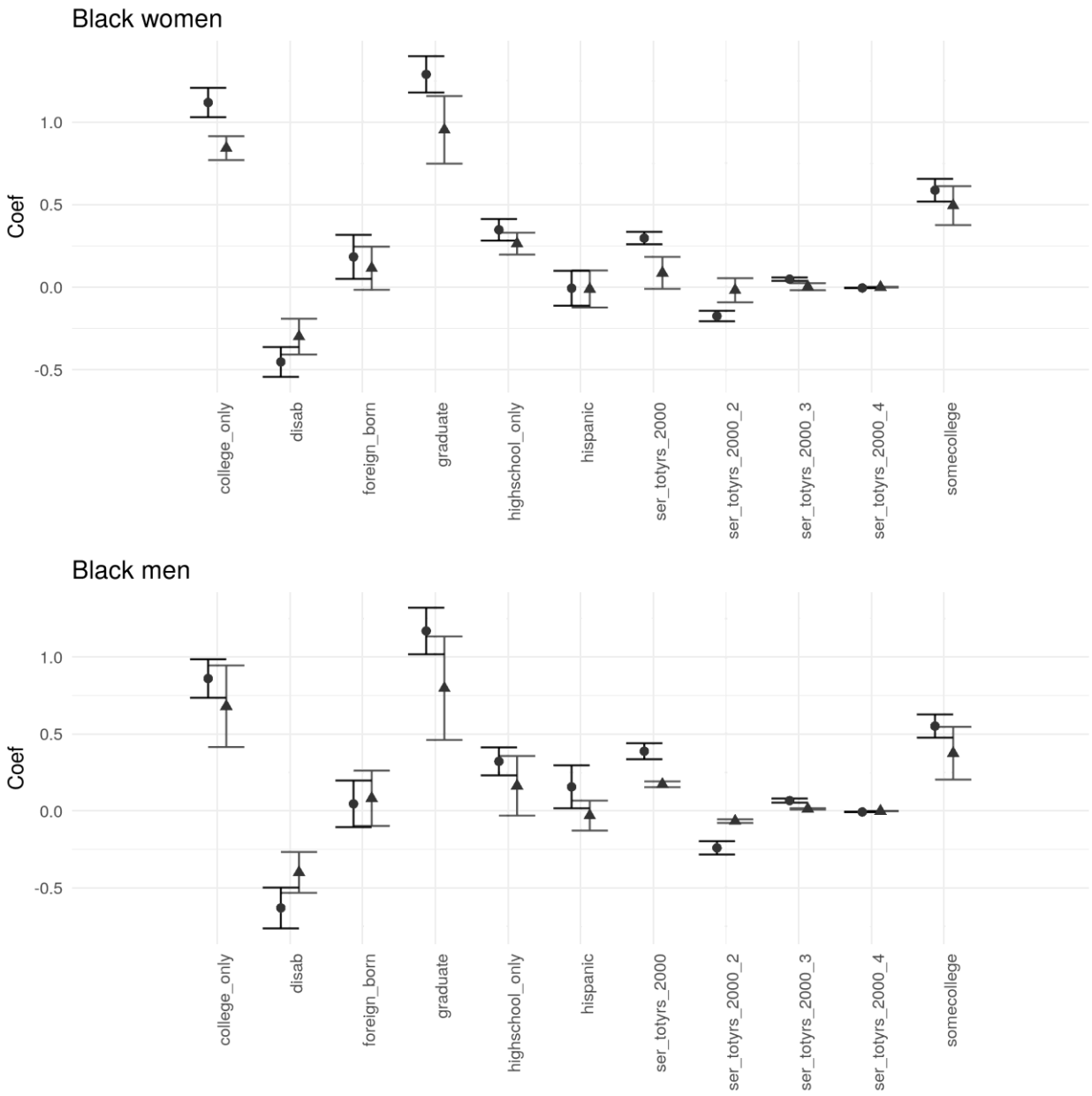

Figure 2: Coefficients of regression of log earnings on various variables, black men and women, year 2000. (Source: Abowd et al. 2006, Table 41, 43)

Sidebox: Practical Synthetic Data Use

The SIPP-SSA-IRS Synthetic Beta File is accessible to users in its current form since 2010. Interested users can request an account by following links at https://www.vrdc.cornell.edu/sds/. Applications are judged solely on feasibility (i.e., the necessary variables are on the SSB). After projects are approved by the Census Bureau, researchers will be given accounts on the Synthetic Data Server. Users can submit validation requests, following certain rules, outlined on the **Census Bureau's website**. Deviations from the guidelines may be

possible with prior approval of the Census Bureau, but are typically only granted if specialized software is needed (other than SAS or Stata), and only if said software also already exists on Census Bureau computing systems. Between 2010 and 2016, over one hundred users requested access to the server, using a succession of continuously improved datasets.

LEHD: Linked establishment and employee records

Data description

The LEHD data links employee wage records extracted from Unemployment Insurance (UI) administrative files from 51 states with establishment-level records from the Quarterly Census of Employment and Wages (also provided by the partner states), the SSA-sourced record of applications for Social Security Numbers ("Numident"), residential addresses derived from IRS-provided individual tax filings, and data from surveys and censuses conducted by the U.S. Census Bureau (2000 and 2010 decennial censuses, as well as microdata from the ACS). Additional information is linked in from the Census Bureau's Employer Business Register and its derivative files. The merged data are subject both to U.S.C. Title 13 and Title 26 protections. For many more details, see (Abowd et al. 2004; Abowd et al. 2009).

From the data, multiple output products are generated. The Quarterly Workforce Indicators (QWI) provide local estimates of a variety of employment and earnings indicators, such as job creation, job destruction, new hires, separations, worker turnover, and monthly earnings, for detailed person and establishment characteristics, such as age, gender, firm age, and firm size (Abowd et al. 2009). The first QWI were released in 2003. The data are used for a variety of analyses and research, emphasizing detailed local data on demographic labor market variables (e.g. Gittings & Schmutte 2016; Abowd & Vilhuber 2012). Based on the same input data, the LEHD Origin-Destination Employment Statistics (LODES) describe the geographic distribution of jobs according to the place of employment and the place of worker residence (Center for Economic Studies 2016). New job-to-job flow statistics measure the movement of jobs and workers across industries and regional labor markets (Hyatt et al. 2014). The microdata underlying these products is heavily used in research, since it provides nearly universal coverage of U.S. workers observed at quarterly frequencies. Snapshots of the statistical production database are made available to researchers regularly (McKinney & Vilhuber 2008; McKinney & Vilhuber 2011; Vilhuber & McKinney 2014).

Disclosure avoidance methods

We describe in detail the disclosure avoidance method used for workplace tabulations in QWI and LODES (Abowd et al. 2012). Not discussed here are the additional disclosure avoidance methods applied in advance of publishing data on job flows (Abowd & McKinney 2016).

Focusing on QWI and LODES is sufficient to highlight the types of confidentiality concerns that arise from working with these linked data, and the kinds of strategies the Census Bureau uses to address them.

In the QWI confidentiality protection scheme, confidential micro-data are considered protected by noise infusion if one of the following conditions holds: (1) any inference regarding the magnitude of a particular respondent's data must differ from the confidential quantity by at least *c%* even if that inference is made by a coalition of respondents with exact knowledge of their own answers (FCSM 2005, p.72), or (2) any inference regarding the magnitude of an item is incorrect with probability no less than *y%*, where *c* and *y* are confidential but generally "large." Condition (1) is intended to prevent, say, a group of firms from "backing out" the total payroll of a specific competitor by combining their private information with the published total. Condition (2) prevents inference of counts of the number of workers or firms that satisfy some condition (say, the number of teenage workers employed in the fast food industry in Hull, GA) assuming item suppression or some additional protection, like synthetic data, when the count is too small.

Complying with these conditions involves the application of statistical disclosure limitation throughout the data production process. It starts with the job-level data that record characteristics of the employment match between a specific individual and a specific workplace, or establishment, at a specific point in time. When the job-level data are aggregated to the establishment level, the QWI system adds statistical noise. This noise is designed to have three important properties. First, every job-level data point is distorted by some minimum amount. Second, for a given workplace, the data are always distorted in the same direction (increased or decreased) and by the same percentage magnitude in every period. Third, when the estimates are aggregated, the distortions added to individual data points tend to cancel out in a manner that preserves the cross-sectional and time-series properties of the data. The chosen distribution is a ramp distribution centered on unity, with a distortion of at least *a%* and at most *b*% (Figure 3).

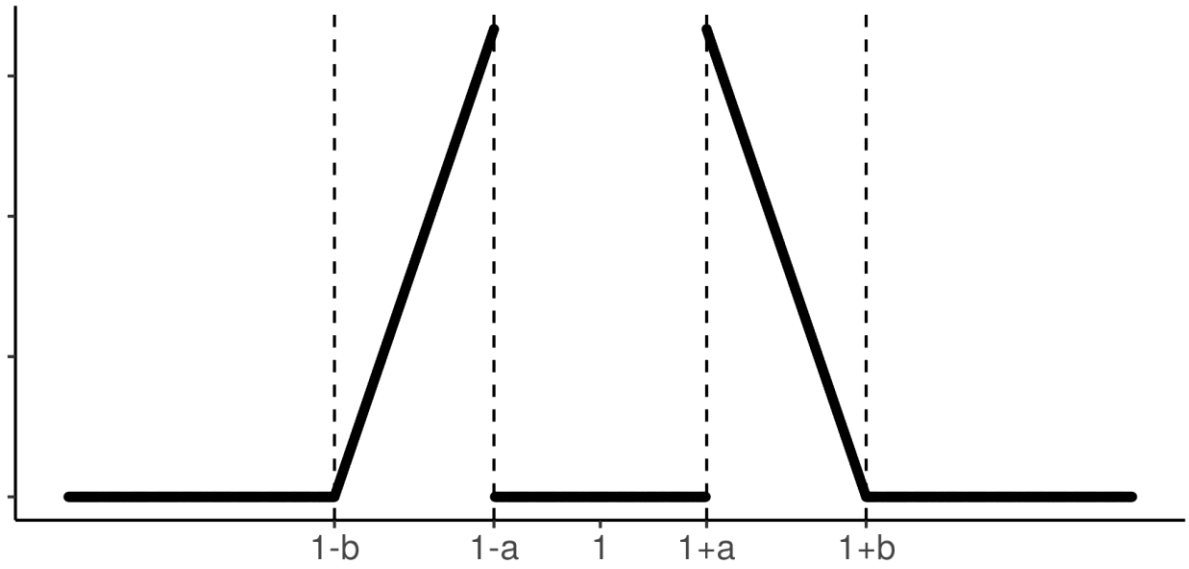

Figure 3: Ramp distribution used in LEHD disclosure avoidance system

All published data from QWI use the same noise-distorted data, and any special tabulations released from the QWI must follow the same procedures. The QWI system extends the idea of multiplicative noise infusion as a cross-sectional confidentiality protection mechanism first proposed by (Evans et al. 1998). A similar noise-infusion process has been used since 2007 to protect the confidentiality of data underlying the Census Bureau's County Business Patterns (Massell & Funk 2007) and was tested for application to the Commodity Flow Survey (Massell et al. 2006).

In addition to noise infusion, the QWI confidentiality protection system uses weighting, which introduces an additional difference between the confidential data item and the released data item. Finally, when a statistic meant to be published turns out to be based on data from fewer than three persons or establishments, it is suppressed. Suppression is only used when the combination of noise infusion and weighting may not distort the publication data with a high enough probability to meet the criteria laid out above; however the suppression rate is much lower than in comparable tabular publications, such as the QCEW.[10] An alternative to suppression (proposed by Gittings 2009; Abowd et al. 2012) uses a synthetic data model that replaces suppressed values with samples drawn from an appropriate posterior predictive distribution. The hybrid system incorporating both noise-infused and synthetic data allows the release of data without suppressions. The confidentiality protection provided by the hybrid system without suppressions is comparable to the protection afforded by the system using the noise infusion system with suppressions, but the analytical validity of the data produced by the hybrid system is improved because the synthetic data are better than the best inference an external user can make regarding the suppressions (Gittings 2009).

[10] Not all estimates are subject to suppression. Estimates such as employment are subject to suppression. Continuous dollar measures like payroll are not (Abowd et al. 2009; Abowd et al. 2012).

The LEHD Origin-Destination Economic Statistics (LODES) provides aggregated information on where workers are employed (Destinations) and where they live (Origins), along with the characteristics of those places. As the name implies, the data are intended for use in understanding commuting patterns and the nature of local labor markets. The fundamental geographic unit in LODES is a Census block, and thus much more detailed than QWI for which data are published as county-level aggregates. LODES is tabulated from the same microdata as the QWI, and for workplaces (the destination), uses a variation of the QWI noise infusion technique. Cells that do not meet the publication criteria of the QWI continue to be suppressed in LODES, but are replaced using synthetic data.[11] For residences (the origin), the protection system relies on a provably-private synthetic data model (Machanavajjhala et al. 2008). A statistical model is built from the data, as the posterior predictive distribution (PPD) of release data *X'* given the confidential data *X*: *Pr[X'|X].* Synthetic data points are sampled from the model *X', and released*. In general, to satisfy *differential privacy* (Dwork 2006; Dwork et al. 2006, 2017), the amount of noise that must be injected into the synthetic data model is quite large, typically rendering the releasable data of low utility. The novelty of the LODES protection system was to introduce the concept of "probabilistic differential privacy," and early variant of what are now called approximate differential privacy systems. By allowing the differential privacy guarantee (parametrized by ε) to fail in certain rare cases (which occur with probability δ), (ϵ, δ)-probabilistic differential privacy (Machanavajjhala et al. 2008) improves the analytical validity of the data greatly. LODES uses Census tract-to-tract relations to estimate the PPD for the block-to-block model. A unique model is estimated for each block, recovering the likelihood of a place of residence conditional on place of work and characteristics of the workers and the workplaces. Several additional measures further improve the privacy and analytical validity of the model (see Machanavajjhala et al. 2008 for further details). The resulting privacy-preserving algorithm guarantees ϵ-differential privacy of 8.99 with 99.999999% confidence ($\delta = 10^{-6}$).

Disclosure avoidance assessment for QWI

The extent of the protection of the QWI micro-data can be measured in two ways: showing the percentage deviation as a measure of the uncertainty about the true value that one can infer from the released value, and the amount of reallocation of small cells (less than 5 entities in a tabulation cell).[12] Each cell underlying the tabulation is for a statistic $X_{kt}$where *k* is a cell defined by a combination of age, gender, industry, and county, and for all released time periods for the states at the time of these experiments.[13] A comparison of the undistorted, unweighted data with

[11] Similar methods have been discussed for the QWI (Abowd et al. 2012; Gittings 2009), but not yet implemented.

[12] The comparisons were computed using custom internal tabulations as well as published numbers, for two states (Illinois and Maryland). Only Maryland is reported here.

[13] The disclosure avoidance assessment was run when first releasing the QWI, in 2003, and are reproduced here as they were presented to the Disclosure Review Board then. At the time, QWI were available for industry classifications according to the Standard Industrial Classification (SIC), 1987 definitions. Modern QWI are available for North American Industry Classification System (NAICS), 2012 definitions. The basic conclusion does not change.

the published data for $X = B$ (**Table 1**) illustrates the combined contribution of weighting, noise infusion, and item suppression.

Table 1: Comparison of unweighted confidential tabulations against published tabulations, QWI

	Published count						
Unweighted count	Suppressed	0	1	2	3	4	5 or more
0	1.06	**98.94**	0	0	0	0	0
1	99.9	0.09	0	0	0	0	0
2	85.71	0.04	0	0	13.9	0.32	0.02
3	23.54	0.03	0	0	**40.18**	33.6	2.65
4	18.06	0.02	0	0	2.22	**33.67**	46.04
5 or more	8.44	0.01	0	0	0.02	0.26	**91.26**
Total number of cells: 4,659,408, Maryland, 2003. Source: Abowd et al, 2012.							

The table entries can be interpreted as the conditional probabilities of publishing the column entry given the confidential row entry. The table is therefore also informative about how much can be learned about the confidential entry from the published data. Table 1 also reports the amount of suppression after weighting and noise-infusion as it relates to the original raw value. All single-individual cells are suppressed. This is not true for two-person cells, some of which have a weighted value that lies above the suppression threshold, in which case the weighted estimate is released. The converse is true for cells with three individuals. Due to weighting, some of these cells have weighted, undistorted values that lie below the suppression threshold, and are consequently suppressed. Cells that contain count data based on fewer than three firms also generate suppressions, which are included in the suppression totals. Overall, at the level of detail analyzed here (SIC3 × county × time × sex × age), around 25% of the beginning of period employment cells are suppressed. For more aggregate tabulations, for instance at the SIC Division level, that percentage falls to between 5% and 10%. Note that there are never any complementary suppressions.

Total payroll, on the other hand, is a dollar magnitude, not an employment count, and is never suppressed. The combination of weighting and distortion is sufficient to protect the confidentiality of this item without suppression because if the item is based on a single person or establishment, then the minimum distortion of the underlying micro-data applies. If the item is based on 2 employers or establishments then both micro-data items have been distorted by at least the minimum percentage. Knowledge of one's own value does not help in inferring another's value because both data items were distorted in an unknown direction by an unknown minimum percentage. Even an accurate inference about one's own distortion factor supplies no

information about the other parties' distortion factor, thus protecting that item by at least the minimum distortion factor in each direction.

Analytical validity assessment for QWI

The noise infusion algorithm for QWI is designed to preserve validity of the data for particular analysis tasks. We demonstrate analytical validity using two statistics: time-series properties of the distorted data relative to the confidential data of several estimates, and the cross-sectional unbiasedness of the published data for beginning-of-quarter employment *B*. The unit of analysis is an interior sub-state geography × industry × age × sex cell *kt*.[14] Analytical validity is obtained when the data display no bias and the additional dispersion due to the confidentiality protection system can be quantified so that statistical inferences can be adjusted to accommodate it.

Time-series properties of distorted data

We estimate an AR(1) for the time series associated with each cell *kt*. For each cell, the error $\Delta r = r - r^*$ is computed, where *r* and $r^*$are the first-order serial correlation coefficient computing using confidential data and protected data, respectively. Table 2 shows the distribution of the errors ∆r across SIC-division × county cells, for accessions *A,* beginning-of-quarter employment *B,* full-quarter employment *F,* net job flows *JF,* and separations *S* (for additional tables, see Abowd et al. 2012). The table shows that the time series properties of the QWI remain largely unaffected by the distortion. The central tendency of the bias (as measured by the median of the ∆r distribution) is never greater than 0.001, and the error distribution is tight: the semi-interquartile range of the distortion for *B* in Table 2 is 0.022, which is less than the precision with which estimated serial correlation coefficients are normally displayed.[15] The overall spread of the distribution is slightly higher when considering two-digit SIC × county and three-digit SIC × county cells (not reported here), due to the greater sparsity. The time series properties of the QWI data are unbiased. The small amount additional noise in the time series statistics is, in general, economically meaningless.

[14] Sub-state geography in all cases is a county, whereas the industry classification is the Standard Industrial Classification (SIC 1987).

[15] The maximum semi-interquartile range for any SIC2-based variables is 0.0241, and for SIC3-based variables, 0.0244.

Table 2: Distribution of errors Δr in first-order serial correlation, QWI

Variable	*Median*	*Semi interquartile range*
Accessions	-0.000542	0.026314
Beginning-of-Quarter Employment	0.000230	0.021775
Full-Quarter Employment	0.000279	0.018830
Net Job Flows	-0.000025	0.002288
Separations	0.000797	0.025539

Cross-sectional unbiasedness of the distorted data

The distribution of the infused noise is symmetric, and allocation of the noise factors is random. The data distribution resulting from the noise infusion should thus be unbiased. We compute the bias ΔX in each cell *kt*, expressed in percentage terms:

$$\Delta X_{kt} = \frac{X^*_{kt} - X_{kt}}{X_{kt}} \times 100$$

Evidence of unbiasedness is provided by Figure 4, which shows the distribution of the bias for *X* =*B.* [16] The distribution of ΔB has most mass around the mode at zero percent. Also, as is to be expected, secondary spikes are present around ±c, the inner bound of the noise distribution.

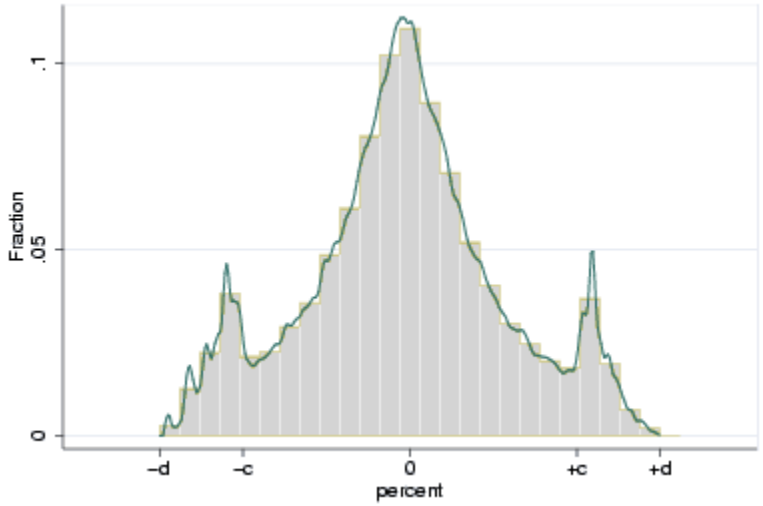

Figure 4: Distribution of ΔB in Maryland. For details, see text.

[16] Data for Maryland. For additional variables and states, see (Abowd et al. 2012). All histograms are weighted by B_{kt}. Industry classification is three-digit SIC (industry groups).

Sidebox: Do-it-yourself noise infusion
The interested user might consult a simple example (with fake data) at https://github.com/labordynamicsinstitute/rampnoise (Vilhuber 2017) that illustrates this mechanism.

Physical and legal protections

The provision of very detailed micro-tabulations or public-use microdata may not be sufficient to inform certain types of research questions. In particular, for business data the thresholds that trigger SDL suppression methods are met far more often than for individuals or households. In those cases, the research community needs controlled access to confidential microdata. Three key reasons why access to microdata may be beneficial are:

(i) "microdata permit policy makers to pose and analyze complex questions. In economics, for example, analysis of aggregate statistics does not give a sufficiently accurate view of the functioning of the economy to allow analysis of the components of productivity growth;
(ii) access to microdata permits analysts to calculate marginal rather than just average effects. For example, microdata enable analysts to do multivariate regressions whereby the marginal impact of specific variables can be isolated;
(iii) broadly speaking, widely available access to microdata enables replication of important research" (United Nations 2007, p.4)

As we've outlined above, many of the concerns about confidentiality have either removed or prevented creation of public-use microdata versions of linked files, exacerbating the necessity of providing alternate access to the confidential microdata.

NSOs and survey organizations usually provide access to confidential linked data within restricted-access data centers. In the United States, this means either using one of 30 secure sites managed by the Census Bureau as part of the Federal Statistical Research Data System (FSRDC),[17] or going to the headquarters of the statistical agency. Similarly, in other countries, access is usually restricted to headquarters of NSOs. Secure enclaves managed by NSOs used to be rare. In the 1990s and early 2000s, an expansion of existing networks and the creation of new, alternate methods of accessing data housed in secure enclaves occurred in several countries. Access methods may be through physical travel, remote submission, or remote

[17] See https://www.census.gov/fsrdc (accessed on December 15, 2017).

processing. However, all methods rely on two fundamental elements. First, the researcher accessing the data is mostly free to choose the modelling strategy of her choice, and is not restricted to the tables or queries that the data curator has used for published statistics. Second, the output from such models is then analyzed to avoid unauthorized disclosure, and subsequently released to the researcher for publication.

Several methods are currently used by NSOs and other data collecting agencies to provide access to confidential data. The following sections will describe each of them in turn.[18]

Statistical data enclaves

Statistical data enclaves, or Research Data Centers, are secure computing facilities that provide researchers with access to confidential microdata, while putting restrictions on the content that can be removed from the facility. The different advisory committees of the two largest professional association (American Statistical Association, ASA, and the American Economic Association, AEA), pushed for easier and broader access for researchers as far back as the 1960s, though the emphasis then was on the avoiding the cost of making special tabulations. The AEA suggested creating Census data centers at selected universities (Kraus 2013). In the 1990s and early 2000s, similar networks started in other countries. In Canada, the Canadian Foundation for Innovation (CFI) awarded a number of grants to open research data centers, with the first opening at McMaster University (Hamilton, Ontario) in 2000.[19] The creation of the RDCs was specifically motivated by the inability to ensure confidentiality while providing usability of longitudinally linked survey data (Currie & Fortin 2015).

In the United States, a 2004 grant by the National Science Foundation laid the groundwork for subsequent expansion of the (then Census) Research Data Center network from eight locations, open since the mid-1990s, to over 30 locations in 2017. One of the key motivations was to make the newly available linked administrative data at LEHD accessible to researchers. The network operates under physical security constraints managed by the Census Bureau and the Internal Revenue Service, in locations that are considered part of the Census Bureau itself, and staffed by Census Bureau employees.

Statistical data enclaves can be central locations, in which a single location at the statistical agency is made available to approved researchers. In the U.S., NCHS and BLS follow this model, in addition to using the FSRDC network. In Canada, business data can be accessed at Statistics Canada headquarters, while other data may be accessed both there and at the geographically dispersed RDCs, which obtain physical copies of the confidential data.

Some facilities are hybrid facilities. The statistical processing occurs at a central location, but the secure remote *access* facilities are distributed geographically. The U.S. FSRDCs have worked this way since the early 2000s. A central computing facility is housed in the Census Bureau's primary data center. Secure remote access is provided to approved researchers at

[18] The section draws on (Weinberg et al. 2007; Vilhuber 2013).
[19] For an extensive history of the Canadian Research Data Center Network, see (Currie & Fortin 2015).

designated sites throughout the county, namely the FSRDCs. Each of the FSRDC sites is a secure Census Bureau facility that is physically located on controlled premises provided by the partner organization, often a university or Federal Reserve Bank. The German IAB locates certified thin clients in dedicated rooms at partner institutions. Secure spaces are costly to build and certify. Recently, institutions in the UK have attempted to reduce the cost by commoditizing such secure spaces (Raab et al. 2015). In France, the Centre d'accès sécurisé distant aux données (CASD) has a secure central computing facility, and allows for remote access through custom secure devices from designated but otherwise ordinary university offices, which satisfy certain physical requirements, but are not dedicated facilities. Similar arrangements are used by Scandinavian NSOs, as well as by survey organizations such as the HRS. Remote access to full desktop environments within the secure data enclave, commonly referred to as “virtual desktop infrastructure” (VDI), from regular laptops or workstations, is increasingly common.

The location of remote access points is often limited to the country of the data provider (United States, Canada), or to countries with reciprocal or common enforcement mechanisms (within the European Union, for European NSOs). Cross-border access, even within the European Union, remains exceedingly rare, with only a handful of cross-border secure remote access points open in the European Union. The most prolific user of cross-border secure remote access points, as of this writing, is the German IAB, with multiple data access points in the United States and a recently opened one in the United Kingdom.

Remote processing

Two other alternative remote access mechanisms are often used: manual and automatic remote processing. Manual remote processing occurs when the remote “processor” is a staff member of the data provider. This can be as simple as sending programs in by email, or finding a co-author who is an employee of the data provider. The U.S. NCHS, German IAB, and Statistics Canada provide this type of access. Generally, the costs of manual remote processing are paid by the users.

More sophisticated mechanisms automate some or all of the data flow. For instance, programs may be executed automatically based on email or web submission, but disclosure review is performed manually. This method is used by the IAB’s JoSuA (Institute for Employment Research 2016). Fully automated mechanisms, such as LISSY (Luxembourg), ANDRE (U.S. NCHS), DAS (U.S. NCES), Australia's Remote Access Data Laboratory (RADL), Canada’s Real Time Remote Access (RTRA), generally restrict the command set from the allowed statistical programming languages (SAS, Stata, SPSS) and limit what the users can do to certain statistical procedures and languages for which known automated disclosure limitation procedures have been implemented.

Most of these systems only provide access to household and person surveys. Of the known systems surveyed above, only Australia's RADL systems and the Bank of Italy's implementation of LISSY (Bruno et al. 2009; Bruno et al. 2014) seem to provide access to *business* microdata through automated remote processing facilities.

Licensing

Users of secure research data centers always sign some form of legally binding user or licensing agreement. These agreements describe acceptable user behavior, such as not copying or photographing screen contents. However, licensing alone may also be used to provide access to restricted-use microdata outside of formal restricted access data centers. In general, the detail in licensed microdata files is greater than in the equivalent (or related) public-use file, and may allow for disclosure of confidential data if inappropriately exploited. For this reason, licensed microdata files tend to have several additional levels of disclosure avoidance methods applied, including output review in some cases. For instance, even without linkages, the HRS licensed files have more detailed geography on respondents (county, say, rather than Census region), but do not have the most detailed geography (GPS coordinates or exact address). Generally, the legally enforceable license imposes restrictions on what can be published by the researchers, and restricts who can access the data, and for what purpose. The contracting organization is the researcher's university, which is subject to penalties such as loss of eligibility status for research grants if the license is violated.

In the United States, some surveys (NCES, NLSY, HRS) use licensing to distribute portions of the data they collect on their respondents. Commercial data providers (COMPUSTAT, etc.) also license the data distributed to researchers. Penalties for license infractions range from restricting future research grant funding, for example in HRS, to monetary penalties, for example in commercial data licenses. We are not aware of any studies that quantify the violation rates or financial penalties actually incurred due to license violations. Licensing may be limited by the enforceability of laws or contracts, and thus may be limited to residents of the same jurisdiction in which the data provider is housed. Often, some licensing is combined with the creation of ad-hoc data enclaves, the simplest of these being stand-alone, non-networked computer workstations.

Disclosure avoidance methods

Data enclaves exist to allow researchers to perform analyses within the restricted environment, and then extract or publish some form of statistical summary that can be released from the secure environment. Generally, these summaries are estimates from a statistical model. In general, model-based output is evaluated according the same criteria traditionally used for tabular output (minimum number of units within a reporting cell, minimum percentage of global activity within a reporting cell). In contrast to licensing arrangements, which allow researchers to self-monitor, statistical data enclaves have regimented output monitoring, typically by staff of the data provider. Generally, released statistical outputs are registered in some fashion, but documentation of the full provenance chain may be limited.

No systematic attempt has been made, to our knowledge, to measure formally the cumulative privacy impact of model-based releases because the science and technology for doing so are too immature. Remote processing facilities, on the other hand, when using automated mechanisms, rely on several practices to reduce the risk of disclosure. First, they limit the scope

of possible analyses to those for which the agency has developed safe procedures. The number of times a researcher may request releases may also be limited. Nevertheless, most agencies recognize that this review system does not scale because the infeasibility of a full accounting of all possible query combinations over time. In general, they apply basic disclosure avoidance techniques such as suppression, perturbation, masking, recoding, and bootstrap sampling of the input data to each project separately. Some systems apply automated analysis of log and output files (Schouten & Cigrang 2003), although often a manual review is also included (O'Keefe et al. 2013). Some systems provide for self-monitored release of model results, either under licensing or remote access. There are also limitations on quantity and frequency of self-released results, combined with sampling by human reviewers. More sophisticated tools, such as perturbation or synthesizing of estimated model parameters, have been proposed (Reiter 2003). Finally, some such systems require review of the draft research paper before submission to any publication medium including online preprint repositories like ArXiv.org.

All three of the examples of linked data provided in this paper rely on some version of secure data enclaves to provide microdata access to approved researchers. HRS data are made available to tenure-track researchers who sign a data use agreement and provide documentation of a secure local computing environment. An additional option for HRS data is to visit to the Michigan Center on the Demography of Aging data enclave, which makes data accessible to researchers in a physical data enclave at "headquarters," like many NSOs. More recently, HRS has started to offer secure VDI access to researchers. The confidential data underlying the SSB, and against which validation requests are run, are also available either within the FSRDC network, or by sending validation requests by email to staff at Census headquarters (a form of "remote processing"). LEHD microdata are only available through the FSRDC.

An open question is whether the disclosure risks addressed through physical security measures are greater for linked data. Enabling researchers to measure some of the heuristic disclosure risk such as *n* cell count or *p*-percent rule (O'Keefe et al. 2013) becomes more important when any possible combination of *k* variables (*k* large) leads to small cells or dominated cells. Even subject matter experts cannot assess these situations *a priori*.

Data silos

One concern with the increasing move to multiple distinct access points for confidential data is the "**siloing**" of data. The critical symptom is a physical separation of files in distinct secure data enclaves. The underlying causes are the incompatible legal restrictions on different data. Typically, these restrictions impose administrative barriers to combining data sources for which linking is technically possible.

Such administrative barriers may also be driven by ethical or confidentiality concerns. The question of consent by survey or census respondents may explicitly prevent the linkage of their survey responses or of their biological specimen with other data. For example, the Canadian Census long form of 2006 offered respondents the option to either answer survey questions on

earnings, or consent to linking in their tax data on earnings. In the 2016 census, the question was no longer asked, and users were simply notified that linkage would happen.

In the case of the LEHD data, as of December 2015, all 50 states as well as the District of Columbia had signed agreements with the Census Bureau to share data and produce public-use statistics. It would thus seem possible for researchers to access a comprehensive LEHD jobs database through the Federal Statistical Research Data Center network, by linking together the job databases from 51 administrative entities. However, all but 12 of the States had declined to automatically extend the right to use the data to external researchers within the FSRDC network. Nevertheless, some of the same states that declined to provide such permission in the FSRDC give access to researchers through their state data centers or other means. The Unemployment Insurance state-level data is thus siloed, and researchers may be faced with non-representative data on the American job market. Several European projects, such as Data without Boundaries (DwB), have investigated cross-national access with elevated expectations but relatively limited success (Schiller & Welpton 2014; Bender & Heining 2011). Increasingly, the U.S. Census Bureau and CASD also host data from other data providers, through collaborative agreements, moving towards a reduction of the siloing of data.

Secure multi-party computing may be one solution to this problem (Sanil et al. 2004; Karr et al. 2005; Karr et al. 2006; Karr et al. 2009). However, implementation of such methods, at least in the domain of the social and medical sciences cooperating with NSOs, is in its infancy (Raab et al. 2015). The typical limitations are the throughput of the secure interconnection between the sources and the requirement of manual model output checking. These limitations drastically slow down any iterative procedure.

Conclusions

The goal of this chapter has been to illustrate how confidentiality protection methods can be and have been applied to linked administrative data. Our examples provide a guide to best-practices for data custodians endeavoring to walk the fine line between making data accessible and protecting individual privacy and confidentiality. Our examples also illustrate different paradigms of protection ranging from the more traditional approach of physical security to more modern formal privacy systems and the provision of synthetic data.

In concluding, we note that from a theoretical perspective, there does not appear to be a clear distinction between the threats to confidentiality in linked data relative to unlinked data, or in survey data relative to administrative data. Richly detailed data pose disclosure risks, whether that richness is inherent in the data design, or comes from linkages of variables from multiple sources. Likewise, there are not special methods to protect confidentiality in linked versus unlinked data. Any data with a network, relational, panel or hierarchical structure poses special challenges to data providers to protect confidentiality while preserving analytical validity. Our example of the QWI shows one way this challenge has been successfully managed in a linked data setting, but the same tools could be effective in application to the QCEW, which uses the same frame, but does not involve worker-firm linkages.

However, from a legal perspective, linking two datasets can change the nature of confidentiality protection in a more practical manner. Any output must conform to the strongest privacy protections required across each of the linked datasets. For example, when the LEHD program links SSA data on individuals to IRS data on firms, any downstream research must comply with the confidentiality demands of all three agencies. Likewise, the data must conform to the U.S. Census Bureau publication thresholds for data involving individuals and firms. Hence, linking data can produce a maze of confidentiality requirements that are difficult to articulate, comply with, and monitor. Harmonizing or standardizing such requirements and practices across data providers, both public and private, and across jurisdictions would be helpful. Privacy and confidentiality issues also invite updated and continuing research on the demand for privacy from citizens and businesses, as well as the social benefit that arises from the dissemination of data.

References

Abowd, J M, B E Stephens, L Vilhuber, F Andersson, K L McKinney, M Roemer, and S D Woodcock. 2009. “The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.” In *Producer Dynamics: New Evidence from Micro Data*, edited by T. Dunne and J B Jensen and M J Roberts. University of Chicago Press.

Abowd, John M, R Kaj Kaj Gittings, Kevin L McKinney, Bryce Stephens, Lars Vilhuber, and Simon D Woodcock. 2012. “Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series.” 12–13. U.S. Census Bureau, Center for Economic Studies. http://dx.doi.org/10.2139/ssrn.2159800.

Abowd, John M, John Haltiwanger, and Julia Lane. 2004. “Integrated Longitudinal Employer-Employee Data for the United States.” *The American Economic Review* 94 (2):224–229.

Abowd, John M., and Kevin L. McKinney. 2016. “Noise Infusion as a Confidentiality Protection Measure for Graph-Based Statistics.” *Statistical Journal of the IAOS* 32 (1):127–35. https://doi.org/10.3233/SJI-160958.

Abowd, John M, and Ian M Schmutte. 2015. “Economic Analysis and Statistical Disclosure Limitation.” *Brookings Papers on Economic Activity* 50 (1):221–267.

Abowd, John M, Martha Stinson, and Gary Benedetto. 2006. “Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project.” 1813/43929. U.S. Census Bureau. http://hdl.handle.net/1813/43929.

Abowd, John M, and Lars Vilhuber. 2012. “Did the Housing Price Bubble Clobber Local Labor Market Job and Worker Flows When It Burst?” *The American Economic Review* 102 (3):589–593. https://doi.org/10.1257/aer.102.3.589.

Australian Bureau of Statistics. 2015. *Media Release - ABS Response to Privacy Impact Assessment*. Australian Bureau of Statistics. http://abs.gov.au/AUSSTATS/abs@.nsf/mediareleasesbyReleaseDate/C9FBD077C2C948AECA257F1E00205BBE?OpenDocument.

Bender, Stefan, and Jörg Heining. 2011. “The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing.” *IASSIST Quarterly / International Association for Social Science Information Service and Technology* 35 (3). http://www.iassistdata.org/iq/issue/35/3.

Browning, Martin, Stephen Jones, and Peter J Kuhn. 1995. “Studies of the Interaction of UI and Welfare Using the COEP Dataset.” LU2-153/224-1995E. Unemployment Insurance Evaluation Series. http://publications.gc.ca/collections/collection_2015/rhdcc-hrsdc/LU2-153-224-1995-eng.pdf.

Bruno, Giuseppe, Leandro D’Aurizio, and Raffaele Tartaglia-Polcini. 2009. “Remote Processing of Firm Microdata at the Bank of Italy.” 36. Bank of Italy. http://dx.doi.org/10.2139/ssrn.1396224.

Bruno, Giuseppe, Leandro D’Aurizio, and Raffaele Tartaglia-Polcini. 2014. “Remote Processing of Business Microdata at the Bank of Italy.” In *Statistical Methods and Applications from a Historical Perspective*, edited by Fabio Crescenzi and Stefania Mignani, 239–249. Studies in Theoretical and Applied Statistics. Springer International Publishing. http://link.springer.com/chapter/10.1007/978-3-319-05552-7_21.

Center for Economic Studies. 2016. “LODES Version 7.” OTM20160223. U.S. Census Bureau. http://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf.

Currie, Raymond, and Sarah Fortin. 2015. *Social Statistics Matter: History of the Canadian Research Data Center Network*. Canadian Research Data Centre Network. http://rdc-cdr.ca/sites/default/files/social-statistics-matter-crdcn-history.pdf.

Dalenius, Tore, and Steven P Reiss. 1982. "Data-Swapping: A Technique for Disclosure Control." *Journal of Statistical Planning and Inference* 6 (1):73–85. https://doi.org/10.1016/0378-3758(82)90058-1.

Deang, Lionel P, and Paul S Davies. 2009. "Access Restrictions and Confidentiality Protections in the Health and Retirement Study." 2009–01. U.S. Social Security Administration. https://www.ssa.gov/policy/docs/rsnotes/rsn2009-01.html.

DeSalvo, Bethany, Frank F Limehouse, and Shawn D Klimek. 2016. "Documenting the Business Register and Related Economic Business Data." Working Papers 16–17. Center for Economic Studies. U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/16-17.html.

Duncan, G.T., Jabine, T.B., de Wolf, V.A. (eds.): Panel on Confidentiality and Data Access, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council and the Social Science Research Council, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* National Academy of Sciences, Washington, DC (1993)

Duncan, G.T., M Elliot, and J J Salazar-González. 2011. *Statistical Confidentiality: Principles and Practice*. Statistics for Social and Behavioral Sciences. New York: Springer-Verlag.

Dwork, Cynthia. 2006. "Differential Privacy." In *Automata, Languages and Programming*, edited by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, 4052:1–12. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/11787006_1.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2016. "Calibrating Noise to Sensitivity in Private Data Analysis." *Journal of Privacy and Confidentiality* 7 (3). http://repository.cmu.edu/jpc/vol7/iss3/2/.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9 (3–4):211–407. https://doi.org/10.1561/0400000042.

Evans, Timothy, Laura Zayatz, and John Slanta. 1998. "Using Noise for Disclosure Limitation of Establishment Tabular Data." *Journal of Official Statistics* 14 (4):537–51.

FCSM. 2005. "Report on Statistical Disclosure Limitation Methodology." 22 (Second version, 2005). {Federal Committee on Statistical Methodology}. https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/04/spwp22.pdf.

Fellegi, I P. 1972. "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association* 67 (337):7–18.

Fellegi, Ivan P, and Alan B Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328):1183–1210. https://doi.org/10.1080/01621459.1969.10501049.

Fienberg, Stephen E.: Confidentiality and disclosure limitation. In: Kempf-Leonard, K. (ed.) Encyclopedia of Social Measurement, pp. 463–469. Elsevier, New York, NY (2005)

Gittings, R Kaj, and Ian M Schmutte. 2016. "Getting Handcuffs on an Octopus: Minimum Wages, Employment, and Turnover." *ILR Review* 69 (5):1133–1170. https://doi.org/10.1177/0019793915623519.

Gittings, Robert. 2009. "Essays In Labor Economics And Synthetic Data Methods." Ph.D., Ithaca, NY, USA: Cornell University. https://ecommons.cornell.edu/handle/1813/14039.

Holan, Scott H, Daniell Toth, Marco A R Ferreira, and Alan F Karr. 2010. “Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality.” *Journal of the American Statistical Association* 105 (490):564–577. https://doi.org/10.1198/jasa.2009.ap08629.

Hyatt, Henry, Erika McEntarfer, Kevin McKinney, Stephen Tibbets, and Doug Walton. 2014. “JOB-TO-JOB (J2J) Flows: New Labor Market Statistics From Linked Employer-Employee Data.” Working Papers 14–34. Center for Economic Studies. U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/14-34.html.

Institute for Employment Research. 2016. *Job Submission Application (JoSuA) at the Research Data Centre of the Federal Employment Agency: User Manual*. https://josua.iab.de/gui/manual.pdf.

Jarmin, Ron S, and Javier Miranda. 2002. “The Longitudinal Business Database.” Working Papers 02–17. Center for Economic Studies. U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/02-17.html.

Karp, Paul. 2016. “Census Controversy Shows ABS ‘Needs to Do Better’, Says Statistical Society.” *The Guardian*, August. http://www.theguardian.com/australia-news/2016/aug/09/census-controversy-shows-abs-needs-to-do-better-says-statistical-society.

Karr, Alan F, Xiaodong Lin, Ashish P Sanil, and Jerome P Reiter. 2005. “Secure Regression on Distributed Databases.” *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 14 (2):263–279. https://doi.org/10.1198/106186005X47714.

———. 2006. “Secure Statistical Analysis of Distributed Databases.” In *Statistical Methods in Counterterrorism*, edited by Alyson G Wilson, Gregory D Wilson, and David H Olwell, 237–261. Springer New York. http://link.springer.com/chapter/10.1007/0-387-35209-0_14.

———. 2009. “Privacy-Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products.” *Journal of Official Statistics* 25 (1):125–138.

Kennickell, A B. 1998. “Multiple Imputation in the Survey of Consumer Finances.” In *Proceedings of the Section on Survey Research*. https://www.federalreserve.gov/econresdata/scf/files/impute98.pdf.

Kinney, Satkartar K, Jerome P Reiter, Arnold P Reznek, Javier Miranda, Ron S Jarmin, and John M Abowd. 2011. “Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database.” *International Statistical Review* 79 (3):362–384. https://doi.org/10.1111/j.1751-5823.2011.00153.x.

Kraus, R. 2013. “Statistical Déjà vu: The National Data Center Proposal of 1965 and Its Descendants.” *Journal of Privacy and Confidentiality*. http://repository.cmu.edu/jpc/vol5/iss1/1/.

Little, Roderick JA. 1993. “Statistical Analysis of Masked Data.” *Journal of Official Statistics* 9 (2):407–26.

Machanavajjhala, Ashwin, Daniel Kifer, John M Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. “Privacy: Theory Meets Practice on the Map.” In *Proceedings of the International Conference on Data Engineering*, 277–286. https://doi.org/10.1109/ICDE.2008.4497436.

Massell, P B, and J M Funk. 2007. “Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding That

Preserves Protection." In *Proceedings of the 2007 FCSM Research Conference*. https://fcsm.sites.usa.gov/files/2014/05/2007FCSM_Massell-IX-B.pdf.
Massell, Paul, Laura Zayatz, and Jeremy Funk. 2006. "Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Luisa Franconi, 304–317. Lecture Notes in Computer Science. Springer Berlin Heidelberg. https://doi.org/10.1007/11930242_26.
McKinney, Kevin L., and Lars Vilhuber. 2011a. "LEHD Infrastructure Files in the Census RDC-Overview." Working Papers 11–43. Center for Economic Studies. U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/11-43.html.
McKinney, Kevin L, and Lars Vilhuber. 2011b. "LEHD Infrastructure Files in the Census RDC: Overview of S2004 Snapshot." Working Papers 11–13. Center for Economic Studies. U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/11-13.html.
National Institute on Aging and the National Institutes of Health. n.d. "Growing Older in America: The Health and Retirement Study." University of Michigan. http://hrsonline.isr.umich.edu/index.php?p=dbook.
O'Keefe, Christine M, Mark Westcott, Adrien Ickowicz, Maree O'Sullivan, and Tim Churches. 2013. "Protecting Confidentiality in Statistical Analysis Outputs from a Virtual Data Centre." Joint UNECE/Eurostat work session on statistical data confidentiality.
Raab, Gillian M, Chris Dibben, and Paul Burton. 2015. "Running an Analysis of Combined Data When the Individual Records Cannot Be Combined: Practical Issues in Secure Computation." Joint UNECE/Eurostat work session on statistical data confidentiality. http://www1.unece.org/stat/platform/display/SDCWS15/Statistical+Data+Confidentiality+Work+Session+Oct+2015+Home.
Reiter, J P. 2003. "Model Diagnostics for Remote-Access Regression Servers." *Statistics and Computing* 13:371–380.
———. 2005. "Significance Tests for Multi-Component Estimands from Multiply Imputed, Synthetic Microdata." *Journal of Statistical Planning and Inference* 131 (2):365–377. https://doi.org/10.1016/j.jspi.2004.02.003.
Reiter, Jerome P, Anna Oganian, and Alan F Karr. 2009. "Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data." *Computational Statistics & Data Analysis* 53 (4):1475–1482. https://doi.org/10.1016/j.csda.2008.10.006.
Reiter, Jerry P. 2004. "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation." *Survey Methodology* 30:235–242.
Rubin, Donald B. 1987. "The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm." *Journal of the American Statistical Association* 82 (398):543–46.
———. 1993. "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2):461–468.
Sanil, Ashish P, Alan F Karr, Xiaodong Lin, and Jerome P Reiter. 2004. "Privacy Preserving Regression Modelling via Distributed Computation." In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 677–682. ACM. https://doi.org/10.1145/1014052.1014139.
Schiller, David, and Richard Welpton. 2014. "Distributing Access to Data, Not Data - Providing

Remote Access to European Microdata." *IASSIST Quarterly* 38 (3). http://www.iassistdata.org/deprecated/iq/issue/38/3.
Schouten, Barry, and Marc Cigrang. 2003. "Remote Access Systems for Statistical Analysis of Microdata." 3004. Statistics Netherlands. https://www.oecd.org/std/37502934.pdf.
Sonnega, Amanda, and David R Weir. 2014. "The Health and Retirement Study: A Public Data Resource for Research on Aging." *Open Health Data* 2 (1):576. https://doi.org/10.5334/ohd.am.
Torra, Vicenç, John M. Abowd, and Josep Domingo-Ferrer. 2006. "Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Luisa Franconi, 4302:233–42. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/11930242_20.
United Nations. 2007. "Managing Statistical Confidentiality and Microdata Access - Principles and Guidelines of Good Practice." United Nations Economic Commission for Europe - Conference of European Statisticians. https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf.
U.S. Census Bureau. 2015. *SIPP Synthetic Beta Version 6.0.2*. Washington, DC and Ithaca, NY, USA. http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/.
Vilhuber, Lars. 2013. "Methods for Protecting the Confidentiality of Firm-Level Data: Issues and Solutions." 19. Labor Dynamics Institute. http://digitalcommons.ilr.cornell.edu/ldi/19/.
———. 2017. *Labordynamicsinstitute/rampnoise: Code for Multiplicative Noise Infusion*. https://doi.org/10.5281/zenodo.1116352.
Vilhuber, Lars, and Kevin McKinney. 2014. "LEHD Infrastructure Files in the Census RDC - Overview." 14–26. Center for Economic Studies, U.S. Census Bureau. http://ideas.repec.org/p/cen/wpaper/14-26.html.
Vilhuber, Lars, Abowd, John M., & Schmutte, Ian M. (2017). "Replication Materials for Disclosure Limitation and Confidentality Protection in Linked Data" (Version V20171214) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1116995
Weinberg, Daniel H, John M Abowd, Philip M Steel, Laura Zayatz, and Sandra K Rowland. 2007. "Access Methods for United States Microdata." 07–25. Center for Economic Studies, U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/07-25.html.

Appendix: Technical Terms and Acronyms

Data

- ACS - American Community Survey, a large survey conducted continuously by the US Census Bureau, on topics such as jobs and occupations, educational attainment, veterans, housing characteristics, and several other topics (https://www.census.gov/programs-surveys/acs/)
- BDS - Business Dynamics Statistics, produced by the U.S. Census Bureau, see www.census.gov/ces/dataproducts/bds/ for more details.

- CBP - County Business Patterns, produced by the U.S. Census Bureau, see www.census.gov/programs-surveys/cbp.html for more details.
- COEP - Canadian Out-of-Employment Panel, a survey initially conducted by McMaster University in Canada, subsequently taken over by the Statistics Canada (Browning et al. 1995)
- COMPUSTAT - a commercial database maintained by Standard and Poor's, with information on companies in the US and around the world (http://www.compustat.com/).
- HRS - Health and Retirement Study, a long-running survey run by the Institute for Social Research at the University of Michigan in the United States on aging in the US population (http://hrsonline.isr.umich.edu/)
- LEHD - Longitudinal Employer Household Program at the U.S. Census Bureau, which links data provided by 51 state administrations to data from federal agencies and surveys (https://lehd.ces.census.gov/)
- LODES - LEHD Origin-Destination Employment Statistics describe the geographic distribution of jobs according to the place of employment and the place of worker residence, in part through the flagship webapp OnTheMap (https://onthemap.ces.census.gov/)
- QWI - Quarterly Workforce Indicators, a set of local statistics of employment and earnings, produced by the Census Bureau's LEHD program (https://lehd.ces.census.gov/data/)
- SIPP - Survey of Income and Program Participation is conducted by the U.S. Census Bureau on topics such as economic well-being, health insurance, and food security (https://www.census.gov/sipp/).
- SSB - the SIPP Synthetic Beta File, also known as “SIPP/SSA/IRS Public Use File”

Other Abbreviations

- ABS - Australian Bureau of Statistics, the Australian NSO (http://abs.gov.au/)
- AEA – American Economic Association (https://www.aeaweb.org)
- ASA - American Statistical Association (https://www.amstat.org)
- BLS - Bureau of Labor Statistics, the NSO in the United States providing data on “labor market activity, working conditions, and price changes in the economy.” (https://bls.gov)
- CASD - Centre d'accès sécurisé distant aux données, the French remote access system to most administrative data files (https://casd.eu)
- Census Bureau - the largest statistical agency in the United States (https://census.gov)
- CMS - Center for Medicare and Medicaid Services administers US government health programs such as Medicare, Medicaid, and others (https://cms.gov/).
- EIA - Energy Information Agency, collecting and disseminating information on energy generation and consumption in the United States (https://eia.gov).
- FICA - Federal Insurance Contribution Act, the law regulating the system of social security benefits in the United States
- IAB - Institute for Employment Research at the German Ministry of Labor (http://iab.de/en/iab-aktuell.aspx).

- FSRDC - Federal Statistical Research Data Centers were originally created as the U.S. Census Bureau Research Data Centers. They provide secure facilities for authorized remote access government restricted-use microdata, and are structured as partnerships between federal statistical agencies and research institutions (https://www.census.gov/fsrdc).
- IRS - Internal Revenue Service handles tax collection for the US government (https://irs.gov)
- NCHS - National Center for Health Statistics, the US NSO charged with collecting and disseminating information on health and well-being (https://www.cdc.gov/nchs/).
- NSO - National statistical offices. Most countries have a single national statistical agency, but some countries (USA, Germany) have multiple statistical agencies.
- OASDI - Old Age, Survivors and Disability Insurance program, the official name for Social Security in the United States
- QCEW - Quarterly Census of Employment and Wages is a program run by the BLS, collecting firm-level reports of employment and wages, and publishing quarterly estimates for about 95% of US jobs (https://www.bls.gov/cew/)
- SER - Summary Earnings Records on SSA data
- SSA - Social Security Administration, administers government-provided retirement, disability, and survivors benefits in the United States (https://ssa.gov)
- SSN - Social Security Number, an identification number in the United States, originally used for management of benefits administered by the Social Security Administration, but since expanded and serving as a quasi-national identifier number.
- UI - Unemployment Insurance, which in the United States are administered by each of the states (and District of Columbia)
- U.S.C - United States Code is the official compilation of laws and regulations in the United States

Concepts

- **Analytical validity**: it exists when, at a minimum, estimands can be estimated without bias and their confidence intervals (or the nominal level of significance for hypothesis tests) can be stated accurately (Rubin 1987). The estimands can be summaries of the univariate distributions of the variables, bivariate measures of association, or multivariate relationships among all variables.
- **Coarsening**: a method for protecting data that involves mapping confidential values into broader categories, e.g. a histogram.
- **Confidentiality**: a “quality or condition accorded to information as an obligation not to transmit […] to unauthorized parties” (Fienberg, 2005, as quoted in Duncan et al, 2011). Confidentiality addresses data already collected, whereas privacy (see below) addresses the right of an individual to consent to the collection of data.
- **Data swapping**: Sensitive data records (usually households) are identified based on *a priori* criteria, and matched to “nearby records”. The values of some or all of the other

variables are swapped, usually the geographic identifiers, thus effectively relocating the records in each other's location.

- **Differential privacy**: a class of formal privacy mechanisms. For instance, ε-differential privacy places an upper bound, parameterized by ε, on the ability of a user to infer from the published output whether any specific data item, or response, was in the original, confidential data (Dwork & Roth 2014).
- **Dirichlet-multinomial distribution:** a family of discrete multivariate probability distributions on a finite support of non-negative integers. The probability vector p of the better-known multinomial distribution is obtained by drawing from a Dirichlet distribution with parameter α.
- **Input noise infusion**: distorting the value of some or all of the inputs before any publication data are built or released.
- **Posterior predictive distribution (PPD)** - in Bayesian statistics, the distribution of all possible values conditional on the observed values.
- **Privacy: "an individual's freedom from excessive intrusion in the quest for information and […] ability to choose [… what …] will be shared or withheld from others" (Duncan et al, 1993, quoted in Duncan et al, 2011). See also confidentiality, above.**
- **Sampling**: as part of SDL, works by only publishing a fractional part of the data.
- **Statistical confidentiality or SDL - Statistical disclosure limitation**: can be viewed as "a body of principles, concepts, and procedures that permit confidentiality to be afforded to data, while still permitting its use of for statistical purposes"(Duncan et al. 2011, p.2).
- **Suppression**: describes the removal of cells from a published table if its publication would pose a high risk of disclosure.